PROTEIN SUBCELLULAR LOCALIZATION PREDICTION THROUGH BOOSTING ASSOCIATIVE CLASSIFICATION

[1]L. A. Khandare , [2]Prof. S. K. Shirgave

[1,2]Compute science, D.Y.Patil college of Engg. Kolhapur, Shivaji University Kolhapur(India)

ABSTRACT

Proteins are composed of linear sequences of smaller molecules called amino acids that accomplish most of the functions of the living cell. The goals in cell biology are to identify the subcellular locations of proteins. A number of different mechanisms have been used for prediction of subcellular locations of proteins. In proposed system we just using protein sequence information. The method divides a protein sequence into short k-mer sequence fragments to generate associationrule. A large number of class association rules are mined from the protein sequence examples that range from the N-terminus to the C-terminus. Then, a boosting algorithm is applied to those rules to build up a final classifier.

Keywords —*Associative Classifier, Association Rules, Bioinformatics (Genome Or Protein) Databases, Pattern Recognition.*

I. INTRODUCTION

Proteins are composed of linear sequences of smaller molecules called amino acids that accomplish most of the functions of the living cell. The goals in cell biology are to identify the subcellular locations of proteins. Proteins perform their appropriate functions only when they are located in the correct subcellular locations. Biochemical experiments are required to determine the subcellular localization of a protein, but experiments are time consuming and high effort requires. So it needs to develop computational methods to predict protein subcellular localization automatically and accurately. Predicting protein subcellular localization has used three types of features, such as N-terminal sorting signals, text annotations from protein databases and amino acid compositions. Some of the methods are based on the existence of targeting signals appearing in N-terminal sequences. Targeting signals reside at a specific part of primary sequence such as the N-terminus or C-terminus. Statistical analysis is required to predict subcellular localization. Several methods are based on text information presented in data base these methods are used the database text annotations information from database to predict subcellular localization.

Most of the technique extracted the information from the entire range of amino acid sequences using data mining techniques. Amino acid compositions contain the sequence of protein information. Frequently occurring sequence fragment are extracted from the training set, and arranged into associative classification rules. Associative Classification (AC) is a classification approach that is a combination of both association rule mining and classification.

II. RELATED WORK

A number of different mechanisms have been used for prediction of subcellular locations of proteins. Most of the technique extracted the information from the entire range of amino acid sequences using data mining techniques [1]. A frequent subsequence is a consecutive series of amino acids that appear in more than a certain number of proteins of a specific class. A frequent pattern has the form *X*X*…, in which each 'X' is a frequent subsequence made of consecutive amino acids. There are two classifiers used to predict the location one based on association rule approach, while the other is based on support vector machines (SVMs). One uses frequent subsequences to construct classification rules for outer membrane proteins and the other uses frequent subsequences as features for a support vector machine (SVM). Association rules are generated using frequent sequential pattern. BCAR system is also based upon amino acid sequences. A primary protein sequence is divided into short sequence fragments, from which associative classification rules are generated using a frequent pattern mining algorithm.BCAR apply boosting algoriyhm to generate smaller number of accurate rules. BCAR is developed a classifier which can handle proteins with single label classification.

To improve the subcellular prediction accuracy the combined Text based and sequence based features are used [2], [3]. SherLoc uses localization predictions from four different sequence-based classifiers and from one text-based classifier, and integrates them to produce an improved prediction of the subcellular localization of the input protein. The four sequence-based classifiers utilize four types of biological features which are n-terminal targeting sequences, amino acid composition, and sequence motifs. SVM classifiers are used to identify the first three types of features. The fourth classifier checks for the presence of sequence motifs in a protein sequence. The text-based classifier classifying proteins use textual representations of the proteins in protein database such as Swiss-Prot. A resource that contains documents related to proteins may be used as the source of the text. The output of each classifier is fed into a final classifier. Y. Liu and Z. Guo[3] used associative classifications (CMAR and CPAR) and multi-class Support Vector Machines for protein subcellular localization prediction. They used a text-based features and shows that CPAR and CMAR achieve very similar accuracy as multi-class SVM with better transparency in classification models.

Eskin and Agicthein [4] with a system that combined protein sequence and text information to create a classifier. Starting with a dataset of proteins of which only a small subset had known locations, they used a text-based classifier increase the number of proteins with an assigned location. Where sequence information is divided into short substrings of length k or k-mer. Then joined text-based SVM classifier is trained by using text and sequences. In data mining a new classification approach, associative classification is used, which is a combination of association rule mining and classification [5]. Association mining is used to discover association rules from databases, while associative classifier is used for categorizing unseen data. They apply the AdaBoost algorithm to an associative classification system. Boosting is used for improving the accuracy of learning algorithm. Boosting is to produce a more accurate prediction rule through combining simple and moderately inaccurate rules.

III. PROBLEM STATEMENT

Predicting protein subcellular localization the conventional classifiers can only handle single label problems as well as suffer from low coverage for test protein. The protein may exist in more than one subcellular location.

So more than one rule needs to be used to predict protein subcellular localization. So that assigns possibly more than one label per location as the result of classification. Hence the goal is to design and implement a multilabel classifier which is capable of dealing with multilabel classification and improves the coverage for test protein.

IV. NEED OF WORK

In Bioinformatics, a protein may reside at, or move between, two or more different subcellular locations. The traditional methods don't take multiple-location proteins into account when predicting protein subcellular localization. Unfortunately, most traditional classifiers can handle single-label problems only. They require to combine several classifiers that use different types of feature. When two or more rules match subcellular locations mostly the best rule (e.g., The one with the highest confidence) is selected and used for prediction while the remaining rules are discarded. For better accuracy and the coverage of test protein it is required to develop new methods for dealing with multilabel classification. In multilabel classification more than one rule is used to assigns more than one subcellular location to each protein.

V. PROPOSED WORK

The block diagram of the proposed work is shown in fig. (1). The proposed framework consists of four basic phases: preprocessing, rule extraction, boosting associative classifiers and analysis.

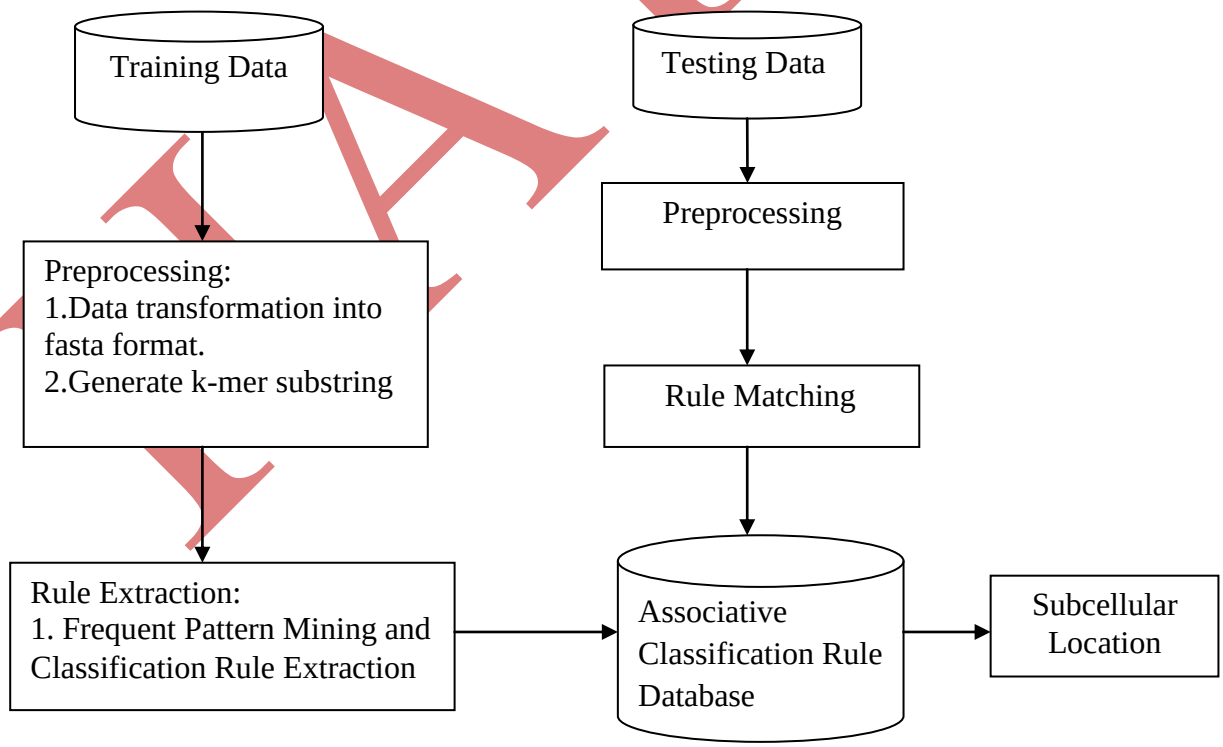

Fig. 1 Block diagram of proposed system

We used two data sets for subcellular localization experiments. The training data sets are Multiloc and TargetP. The TargetP data set contains 3,678 redundancy-reduced protein sequences from the Swiss-Prot database. Plant proteins are annotated with four location labels: chloroplast, mitochondrion, secretory pathway, and other. Nonplant proteins are annotated with three locations: mitochondrion, secretory pathway, and other. The other category consists of cytoplasmic and nuclear locations.

The MultiLoc data set has 5,959 proteins contains animal, fungal, and plant protein sequences from the Swiss-Prot release 42. The plant proteins have 10 subcellular locations: chloroplast, cytoplasm (cyt), endoplasmic reticulum (er), extracellular space (ext), Golgi apparatus (gol), mitochondrion, nucleus (nuc), peroxisome (per), plasma membrane (pm), and vacuole (vac). Animal proteins lack chloroplasts, and have lysosomes (lys) instead of vacuoles.

5.1 Preprocessing

Preprocessing consists of removal of noise and irrelevant data, and in addition to this it also converts into document format which is the required format for rule generation. Preprocessing consists of two step Data transformation into FASTA format and generate k-mer substring. The SWISS-PROT protein sequence data bank is composed of sequence entries. Protein data set is the combination of protein sequence and all known relevant information about a particular protein. It contains details about every protein sequence such as Identification, Accession number, Date, Gene name, Comments, Sequence header etc. So all relevant information about a protein and homologous amino acid sequences are removed. Protein sequences with subcellular locations only keep into the data base.

The protein data bank contains protein sequences, i.e. Contiguous sequence of characters. An example of protein sequence data lines is shown in fig.(2)

GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTASANKNKGIIWGEDTLME
YLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKA

Fig.2 Protein Sequence

Protein sequences are divided into fixed-length sequence fragments (k-mer) with subcellular location labels annotated. ARFF files are generated from fixed-length sequence fragments. Where value of k=3 for better database coverage and accuracy. Protein sequence after fragmentation is shown in Fig. (3).

GDV EKG KKI FIM KCS QCH TVE KGG KHK TGP NLH GLF GRK TGQ APG YSY TAS

ANK NKG IIW GED TLM EYL ENP KKY IPG TKM IFV GIK KKE ERA DLI AYL KKA

Fig.3 k-mer substring

5.2 Rule Extraction

Pattern Generation phase incorporates to find frequent itemsets from a transaction data set. Once frequent itemsets are obtained, it is used to generate association rules with a user specified minimum confidence. Huge numbers of class association rules are generated to improve the coverage for test sequences. The association rule mining generates a set of rules in the form of condition ⇒ class, where the condition is a set of short sequence fragment items and class is label associated with that rule. Where association rule discovery describes correlations between items in a transactional database. The support of a rule is the number of training examples in which the pattern and the class label occur. The confidence of the rule is defined as con f(r_i)/sup(the condition of r_i). Rules are generated based on user-defined minimum support and minimum confidence, which are the parameters used in association rule mining. Generated rules are used by a classification system to predict the class of new items. Weka's implementation of the Apriori association rule finding algorithm is used to generate association rules.

ARFF file is generated by dividing dataset into fixed-length sequence fragments with subcellular location labels. ARFF file is given as an input for Apriori algorithm to generate Association rules.

5.3 Boosting Class Association Rules and Associative Classification

5.3.1 Boosting Class Association Rules

Frequent pattern mining method generates a large number of classification rules. So that we need to filter out some useless or redundant rules and select a small number of high quality and accurate rules, and include in the rule database. So the Adaptive boost (AdaBoost) algorithm is used for boosting class association rules. In Boosting process weights is assigned to each rule, which improves or boosts the prediction ability of individual weak predictors. The algorithm progress iteratively and in each iteration the classifier is improved. While boosting reduces the number of classification rules, it improves performance greatly.

5.3.2 Associative Classification

The task of classification is to build a classifier to predict class labels of unknown objects with high accuracy. A test protein with unknown subcellular localization is converted into a document format same as

training examples. Labels to test example is assigned based on rule obtained during the association rule mining process. Where labels to test example is depend on the confidence of the rule. With existing associative classification techniques, only one class label is associated with each rule derived, and thus the rules are not suitable for the prediction of multiple labels. However, multi-label classification is very useful in protein subcellular localization. Multi-label classification for protein subcellular localization ranking is used to deal with multi-label classification. Rules generated during single-label rule mining are used for label ranking. In ranking, the order set of labels L, so that the topmost t labels are more related to the new object. Then apply thresholding techniques to choose more top three rules matching a new and unknown object during classification.

5.4 Analysis

In this step, experiments are performed to evaluate the performances of the associative classification systems. The classification performances are evaluated by their classification accuracies i.e .the prediction result of each subcellular location. Five-fold cross validation is used, in which a data set is divided into five subsets, four subsets for training and one for testing. The predicted result of each subcellular location is evaluated with specificity (Spec) and sensitivity (Sens) measures.

$$Spec=\frac{TP}{TP+FP}$$

$$Sens=\frac{TP}{TP+FN}$$

The overall performance on a whole data set is evaluated using accuracy, which is defined as

$$Acc=\frac{\sum_{j=1}^{j=M} TP_j}{\sum_{j=1}^{j=M}(TP_j+FN_j)}$$

Where

j is location index, and M is the number of total locations.

VI. CONCLUSION

In previous work Predicting protein subcellular localization they only handle single label problems as well as suffer from low coverage for test protein. The protein may exist in more than one subcellular location. So more than one rule needs to be used to predict protein subcellular localization. to overcome the above problem, this paper proposed rule mining algorithm which support to overcome the existing problem and assign the more than one label to the protein. Hence the goal is to design and implement a multilabel classifier which is capable of dealing with multilabel classification and improves the coverage for test protein.

The efficiency of the proposed method will be evaluated by computing the recall and precision values as well as overall accuracy .the results of system will be compared based upon bench mark dataset for finding better approach.

REFERENCES

[1] Rong She, Fei Chen, Ke Wang, Martin Ester, *Frequent-Subsequence-Based Prediction of Outer Membrane Proteins, SIGKDD,03, August 24-27, 2003.*

[2] H. Shatkay, A. Hoglund, S. Brady, T. Blum, P. Donnes, and O. Kohlbacher, *Sherloc: High-Accuracy Prediction of Protein Localization by Integrating Text and Protein Sequence Data, Bioinformatics, vol. 23, no. 11, pp. 1410-1417, 2007.*

[3] Yifeng Liu , Zhaochen Guo, Xiaodi Ke, Osmar R. Zaïane , *Protein Subcellular Localization Prediction with Associative Classification and Multi-class SVM, ACM-BCB '11, August 2011.*

[4] E. Eskin and E. Agichtein, *Combining Text Mining and Sequence Analysis to Discover Protein Functional Regions, Proc. Pacific Symp. Biocomputing, pp. 288- 299, 2004.*

[5] Yanmin Sun, Yang Wang, "Boosting an Associative Classifier, *IEEE transactions on knowledge and data engineering, VOL. 18, NO. 7, JULY 2006*

[6] Yongwook Yoon and Gary Geunbae Lee, *Subcellular Localization Prediction through Boosting Association Rules, Ieee/Acm Transactions On Computational Biology And Bioinformatics, Vol. 9, No. 2, March/April 2012*

[7] Amos Bairoch, Rolf Apweiler,*The Swiss-Prot Protein Sequence Database User Manual.*